Title: R & R – Residuals and Regression

Brief Overview:

Students will experiment with different values of the slope of a regression line to find a "best line" using the concept of minimizing the sum of the squares of the residuals. They then will use a calculator to find the best-fit line and verify that it produces a least sum of squares of the residuals.

Links to NCTM Standards:

- **Mathematics as Reasoning**
  Students will use the sum of the squares of the residuals as a criterion for choosing the best regression line for a given set of data.

- **Statistics**
  Students will learn how to calculate residuals and use the "least squares" numbers to determine the best-fit line to data.

Grade/Level:

Grades 9-12 (Algebra II, Pre -Calculus, Statistics)

Duration/Length:

This activity will take 2 or 3 hours of class time.

Prerequisite Knowledge:

The student should be able to do the following:

- Perform operations on a TI-83 graphing calculator
- Write the point-slope form of a linear function
- Evaluate a function for different values of the domain
- Work with f(x) function notation

Objectives:

Students will:

- understand what residuals are and how to use them to identify the best-fit line.
- learn to use the TI-83 to explore mathematical outcomes.
- work cooperatively with a partner.
- learn what is meant by "the least squares" criterion.

Materials/Resources

- TI-83 Calculator
- R & R Worksheet

Development/Procedures

1. Group the students in pairs.
2. Teacher presentation of definition of residuals.
3. Teacher instruction on use of LISTS on the TI-83.
4. Complete the R & R worksheet.

Evaluation:

The teacher will evaluate student performance through his/her participation in classroom discussion and by evaluation of written answers to questions on the worksheet.

Extension/Follow Up:

The motivated student could demonstrate that the sum of the residuals is a quadratic function of the slope and that the optimum slope occurs at the vertex of the graph of this function.

Authors:

Betsy K. Bennett	Robert G. Davies	Francis B. Imbrescia
St. Alban's School	Woodberry Forest School	Washington-Lee H.S.
Washington, DC	Woodberry Forest, VA	Arlington, VA

R & R - Residuals and Regression Worksheet

Frank and Gigi have been given the following data points and are asked to find a linear function which best models the data. The (x,y) data points are: (0,2), (1,3), (2,5), (3,6), (4,8), and (5,9). Frank thinks that a good equation is $f(x) = 2x + 1$ while Gigi thinks that $g(x) = x + 3$ is a better equation. Which equation is a better model?

ACTIVITY 1

Fill out the blanks in the table following the directions given below:

1	2	3	4	5	6	7	8
x	y	f(x)	f(x) residuals	f(x) res. squared	g(x)	g(x) residuals	g(x) res. squared
0	2						
1	3						
2	5						
3	6						
4	8						
5	9						

1. Find the values of f(x) and g(x) for the values of x given in column 1.

2. The residuals for a given function are defined to be the difference between the actual values of y and the predicted values of y. For example, consider the function f(x). The residual for x=0 is 2 - f(0) = 2 - [2(0)+1] = 1. For the remainder of the data points, calculate the residuals for Frank's function.

3. Find the residuals for Gigi's function, g(x).

4. Because some of the residuals are positive and some are negative, we square each of the residuals. Do this for each of the residuals for f(x) and for g(x). Enter these numbers in columns 5 and 8 respectively.

5. To judge how well our line fits the data we use the sum of the squares of the residuals. Find the sums of the numbers in columns 5 and 8 respectively.
   Write the sum of column 5 _____. Write the sum of column 8 _____.
   Based on these data, which function [f(x) or g(x)] gives the best fit?_____
   Explain why you chose the function you did.

ACTIVITY 2

Your graphing calculator provides an easy way to graph the data points and to graph both f(x) and g(x).

1. Enter the data points from the table above into your calculator by entering the x-values into L_1 and the y-values into L_2.

2. Now enter f(x) into Y1. Enter g(x) into Y2

3. To graph the data points we must use STAT PLOT and select the scatter plot type of graph. The x-list is L_1, while the y-list is L_2.

4. The calculator will choose the best graphing window if you choose <ZOOM><9>

5. Look at the graph. Which one of the lines is the graph of f(x)? _____

   Which of one of the lines is the graph of g(x)? _____

   Which line is the better model? How can you tell? _____

6. Perhaps you can find a linear function which better fits the data. Enter your equation into Y3 and see whether your line is better than Frank's or Gigi's.

ACTIVITY 3

Earlier you should have discovered that Gigi's function, g(x) = x + 3 was a better model for the data than f(x) = 2x + 1 because the sum of the squares of the residuals for g(x) was less than those for f(x). But is Gigi's function the best linear function we can find, or is there an even better function? Can we find a function which will give an even smaller sum of squares of residuals? In this activity we will try to find a linear function which is better than g(x).

1. Mathematicians know that the line which best fits a set of data points contains the point whose x-coordinate is the average of all the x's and whose y-coordinate is the average of the y's. The coordinates of this point are $(\bar{x}, \bar{y})$, where $\bar{x}$ and $\bar{y}$ are the averages of x and y respectively. From the table, find $\bar{x}$ and $\bar{y}$, either by hand or by using the statistics menu in your calculator. What are the values for $\bar{x}$ and $\bar{y}$?

   $\bar{x}$ = _____ $\bar{y}$ = _____

2. The point-slope form of the equation of a line which contains a point $(\bar{x}, \bar{y})$ and has slope m is

$$y - \bar{y} = m(x - \bar{x}).$$

Solving for y we get

$$y = m(x - \bar{x}) + \bar{y}.$$

Let's put this equation into Y1 so that we can graph lines with different slopes. The calculator can graph several different lines on the same axes. We will graph several different lines with varying slopes at the same time by putting the desired values into a list called SLOPE. Start by trying 1,2, and 3 for the slope. From the home screen put these three numbers into a list named SLOPE. Remember that when you want to change this list, you access SLOPE through the list menu.

3. Delete Y1 and Y2 from the calculator and type in a new Y1. The equation that you type in for Y1 should read:

$$Y1 = SLOPE * (x - "\bar{x}") + "\bar{y}"$$

where "$\bar{x}$" and "$\bar{y}$" are the values that you obtained earlier. You should use at least 6 decimal places.

4. Now graph. You will see three lines with slopes 1,2, and 3. Which is which? Which line provides the best fit? Which line provides the worst?

5. Experiment with different values of the slope expressed to the nearest 0.1 to determine visually which slope provides the best fit. Try, for example, slopes of 1.2, 1.4, and 1.6. Write the equation of the line which you believe to be the best fit.

Equation of "best line": _____

ACTIVITY 4

It should be obvious to you by now that one cannot choose the best fit line by "eyeballing" it. We will use the criteria of least squares to determine the best fit line. We must calculate the sum of the squares of the residuals for each value of m that we wish to test. We will set up our calculators so that for every value of m(the slope) that we wish to try, the calculator will calculate the sum of the squares of the residuals.

1. We will now use variable m to represent the slope. Re-enter your equation into Y1. It should now read

$$Y1 = m * (x - \bar{x}) + \bar{y}$$

where again, your values for $\bar{x}$ and $\bar{y}$ are used.

2. List L_3 will contain the predicted values of y. Go to the top of the list L_3 in the STAT EDIT mode and type "Y1(L_1)". The quotation marks are important and will be explained later.

3. List L_4 will contain the residuals which are the differences between the actual values of y and the predicted values of y. Go to the top of list L_4 in the STAT EDIT mode and type "$L_2 - L_3$".

4. List L_5 will contain the squares of each of the residuals. Go to the top of list L_5 in the STAT EDIT mode and type "$L_4 * L_4$".

5. In order to find the sum of the squares of the residuals (which are in L_5) we must find the sum of the numbers in this list. This will be done manually from the home screen by using the keystrokes

   <2ND> <LIST> <MATH> <SUM> <2ND> <L_5>.

6. The quotation marks used above guarantee that when we change the value of m, the numbers in each list will be recalculated so that we can find the sum of the squares for the line with the given slope. Using your calculator find the sum of the squares of the residuals for several different values of the slope. (Store your value for the slope in m and ask for the sum of L_5. Remember, you should be trying values of m which make the sum of the squares small. Record your results in the table below:

Slope, m									
Sum of Residuals Squared									

7. Write the equation which you believe gives the best fit to the data.

   _____

ACTIVITY 5 (THIS DISCUSSION IS THE CULMINATING ACTIVITY)

You should now understand the process of determining the equation of the best-fit line for a given set of data. Fortunately, every time we wish to determine the best regression line, we do not need to go through the above procedure. The calculator does this for us.
With the data still in L_1 and L_2 have the calculator find the best regression line by using the commands <STAT> <CALC> <LinReg>.

Write the equation of the best regression line: _____

Show, by direct substitution, that $(\bar{x}, \bar{y})$ lies on the best-fit line.

Use your calculator to find the sum of the squares of the residuals for the best fit line determined above.

Sum of squares: _____

Is this sum lower than the lowest sum you found in Activity 4? _____

Explain, in detail, the relationship between the least squares regression line and the sum of the residuals squared for that line.